Modern Database Systems

Introduction to the world of Big Data

Doc. RNDr. Irena Holubova, Ph.D.

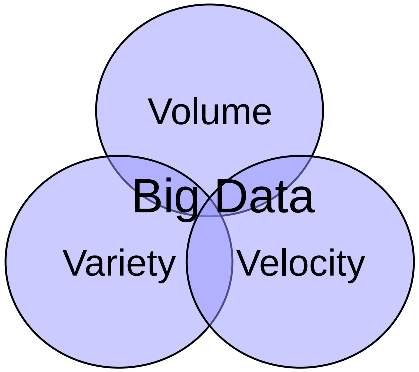
Irena.Holubova@matfyz.cuni.cz

What is Big Data?

- buzzword?
- bubble?
- gold rush?
- revolution?

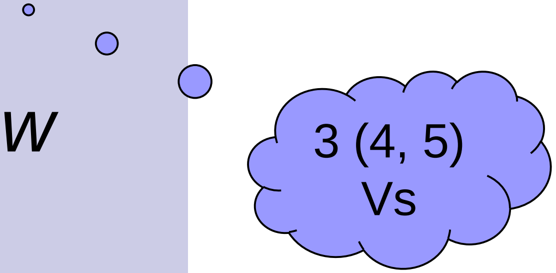
“Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.”
Dan Ariely

What is Big Data?



- No standard definition
- First occurrence of the term: High Performance Computing (HPC)

Gartner: *“**Big Data**” is high **v**olume, high **v**elocity, and/or high **v**ariety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.*



http://www.gartner.com/

Who is Gartner ?

- Information technology research and advisory company
- Founded in 1979 by Gideon Gartner
- HQ in Stanford, Connecticut, USA
  - > 5,300 employees
  - > 12,400 client organizations
- Provides: competitive analysis reports, industry overviews, market trend data, product evaluation reports, …

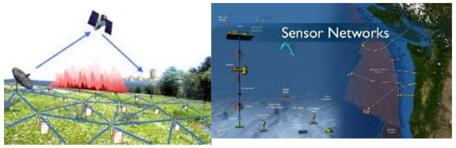
http://www.ibmbigdatahub.com/

What is Big Data?

Mobile devices
(tracking all objects all the time)

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

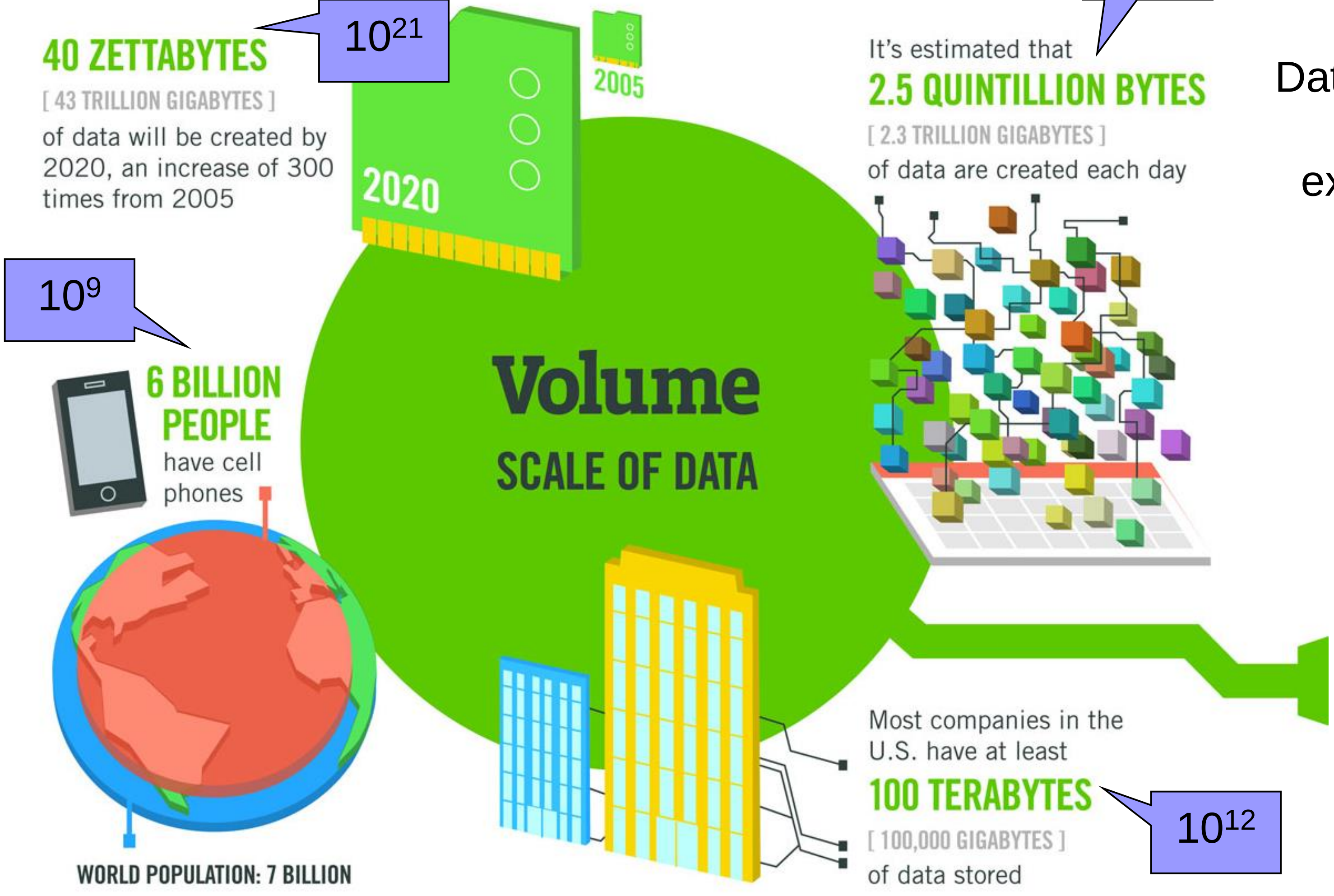


Sensor technology and networks
(measuring all kinds of data)

IBM: *Depending on the industry and organization, **Big Data** encompasses information from internal and external sources such as transactions, social media, enterprise content, sensors, and mobile devices.*
Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

http://www.ibmbigdatahub.com/

Big Data Characteristics: Volume (Scale)



10^{21}

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

10^{18}

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

10^{9}

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Volume

SCALE OF DATA

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

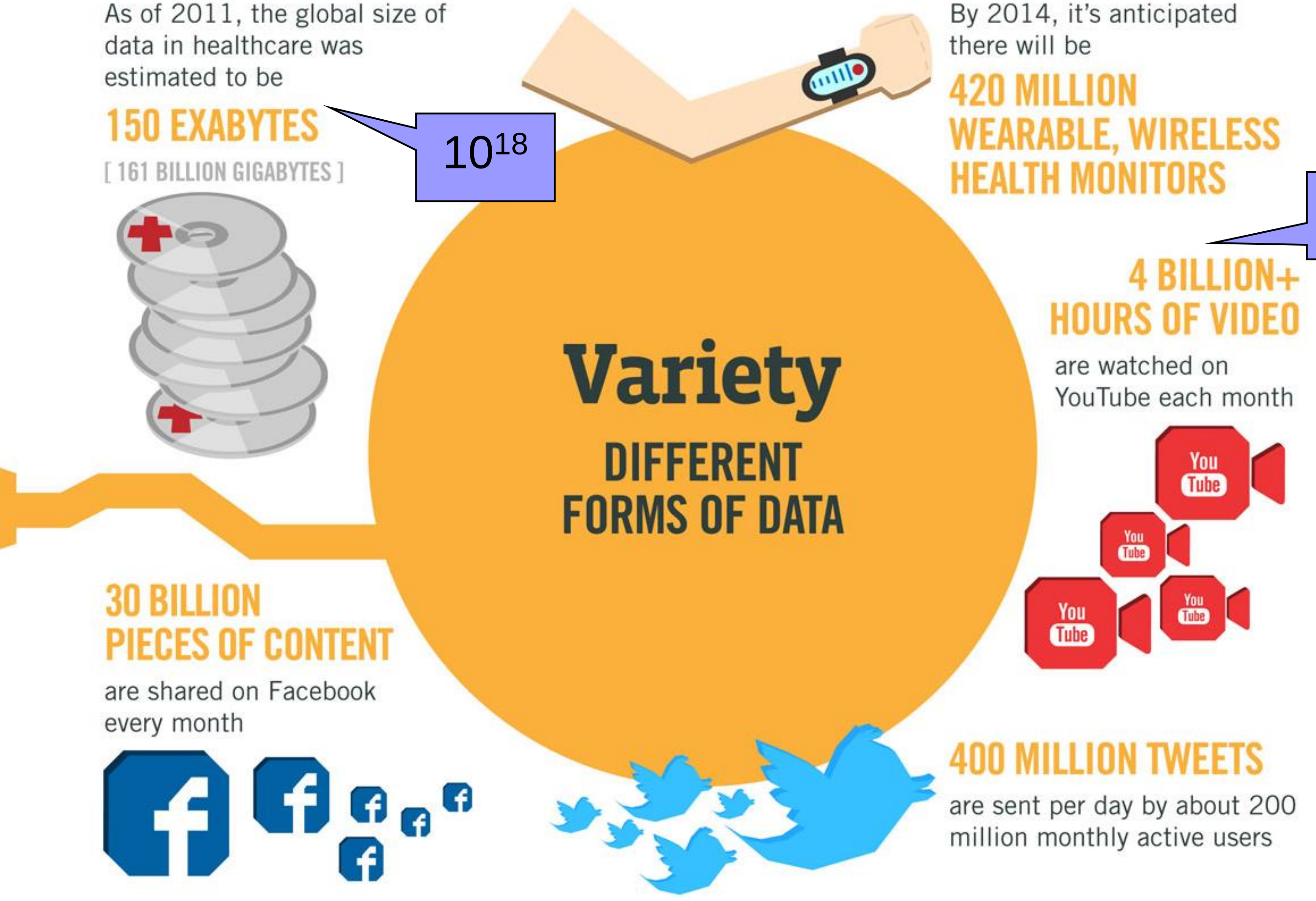
of data stored

10^{12}

Data volume is increasing exponentially, not linearly

http://www.ibmbigdatahub.com/

Big Data Characteristics: Variety (Complexity)


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

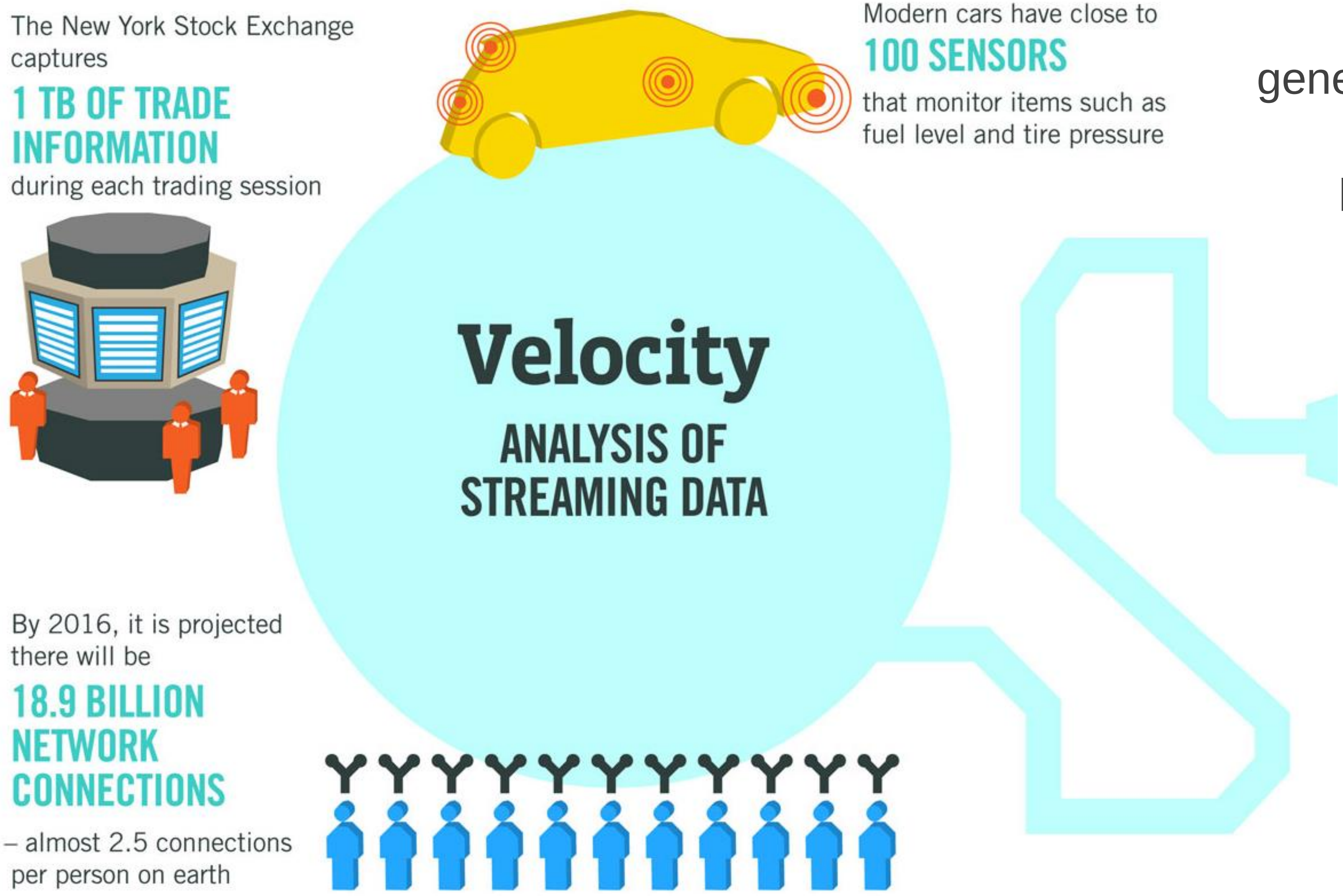
10^{18}

10^{9}

Various formats, types, and structures (from semi-structured XML to unstructured multimedia)

http://www.ibmbigdatahub.com/

Big Data Characteristics: Velocity (Speed)



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

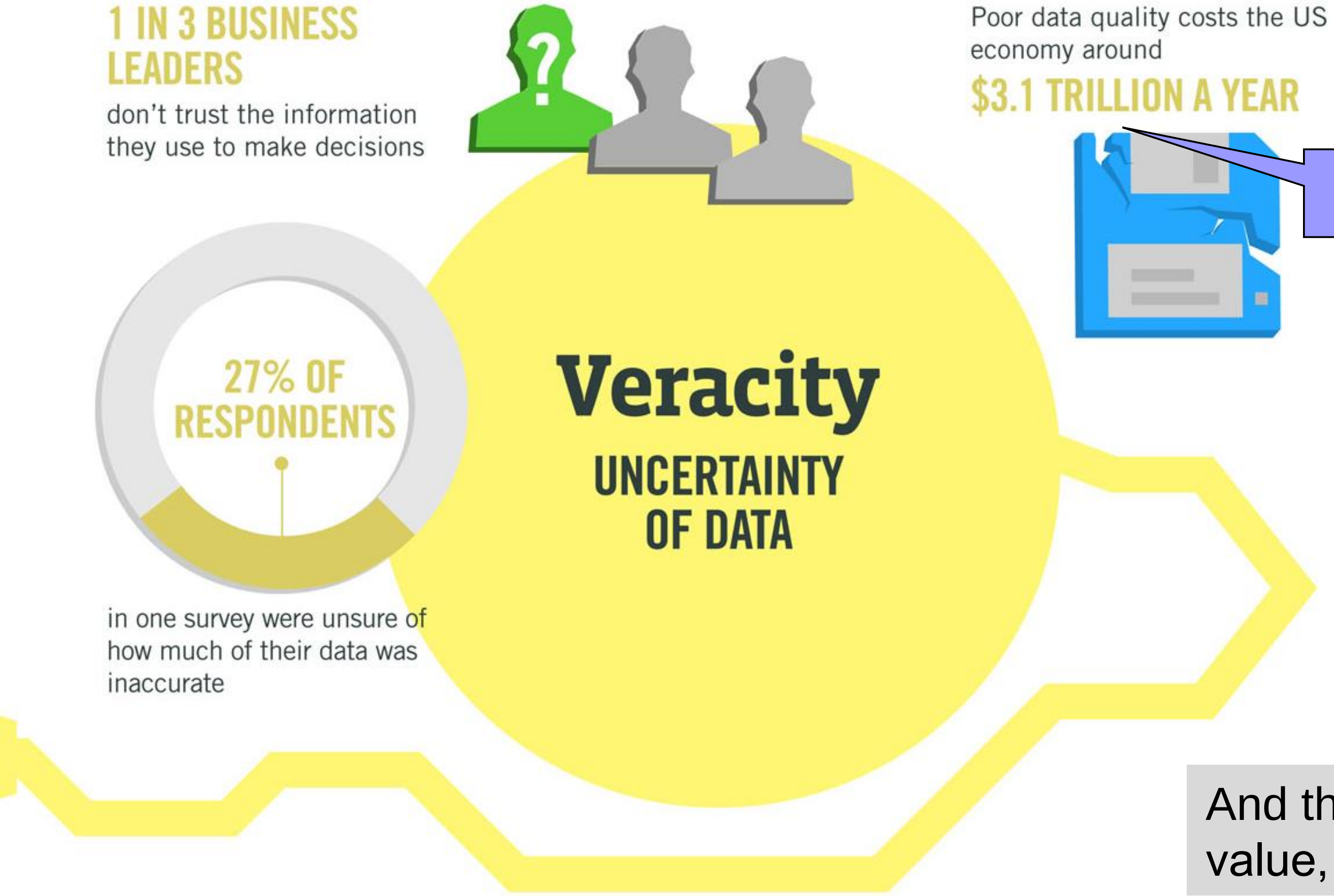
– almost 2.5 connections per person on earth

Data is being generated fast and need to be processed fast

Online Data Analytics

http://www.ibmbigdatahub.com/

Big Data Characteristics: Veracity (Uncertainty)



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around $3.1 TRILLION A YEAR

10^{12}

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

Uncertainty due to inconsistency, incompleteness, latency, ambiguities, or approximations.

And there are new V-s like value, validity, volatility…

Processing Big Data

- OLTP: Online Transaction Processing (DBMSs)
  - Database applications
  - Storing, querying, multiuser access
- OLAP: Online Analytical Processing (Data Warehousing)
  - Answer multi-dimensional analytical queries
  - Financial/marketing reporting, budgeting, forecasting, …
- RTAP: Real-Time Analytic Processing (Big Data Architecture & Technology)
  - Data gathered & processed in a real-time
    - Streaming fashion
  - Real-time data queried and presented in an online fashion
  - Real-time and history data combined and mined interactively

Key Big Data-Related Technologies

- Distributed file systems
- Distributed databases
- Grid computing, cloud computing
- MapReduce and other new paradigms
- Large scale machine learning

Relational Database Management Systems (RDMBSs)

- Predominant technology for storing structured data
  - Web and business applications
- Relational calculus, SQL
- Often thought of as the only alternative for data storage
  - Persistence, concurrency control, integration mechanism, …
- Alternatives: Object databases or XML stores
  - Never gained the same adoption and market share

Modern Database Systems for Specifics of Big Data

- NoSQL databases
  - Key/value, column, document
  - Graph
- NewSQL databases
- Multi-model databases
- Array databases
- …

„NoSQL“

- 1998 first used for a relational database that omitted the use of SQL
  - Carlo Strozzi
- 2009 used for conferences of advocates of non-relational databases
  - Eric Evans
    - Blogger, developer at Rackspace

NoSQL movement = “the whole point of seeking alternatives is that you need to solve a problem that relational databases are a bad fit for”

„NoSQL“

- Not „no to SQL“
  - Another option, not the only one
- Not „not only SQL“
  - Oracle DB or PostgreSQL would fit the definition
- „Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable. The original intention has been modern web-scale databases. Often more characteristics apply as: schema-free, easy replication support, simple API, eventually consistent (BASE, not ACID), a huge data amount, and more“

The End of Relational Databases?

- Relational databases are <u>not</u> going away
- Compelling arguments for most projects
  - Familiarity, stability, feature set, and available support
- We should see relational databases as one option for data storage
  - Polyglot persistence – using different data stores in different circumstances
  - Search for optimal storage for a particular application
    - Multi-model databases

Motivation for NoSQL Databases

- Huge amounts of data are now handled in real-time
- Both data and use cases are getting more and more dynamic
- Social networks (relying on graph data) have gained impressive momentum
    - Special type of NoSQL databases: graph databases
- Full-text has always been treated shabbily by RDBMS

http://royal.pingdom.com/2010/06/18/the-software-behind-facebook/

Example: FaceBook

Statistics from 2010

- 500 million users
- 570 billion page views per month
- 3 billion photos uploaded per month
- 1.2 million photos served per second
- 25 billion pieces of content (updates, comments) shared every month
- 50 million server-side operations per second

2008: 10,000 servers
2009: 30,000 servers
…

→ One RDBMS may not be enough to keep this going on!

Example: FaceBook

Architecture from 2010

Cassandra

- NoSQL distributed storage system with no single point of failure
- For inbox searching

Hadoop/Hive

- An open source MapReduce implementation
- Enables to perform calculations on massive amounts of data
- Hive enables to use SQL queries against Hadoop

Example: FaceBook

Architecture from 2010 and later

Memcached

- Distributed memory caching system
- Caching layer between the web servers and MySQL servers
  - Since database access is relatively slow

HBase

- Hadoop database, used for e-mails, instant messaging and SMS
- Has recently replaced MySQL, Cassandra and few others
- Built on Google's BigTable model
  - Column database

63 Facebook Statistics You Need to Know in 2022

(last update: January 4, 2022)

- 2.91 billion monthly active users
- Facebook has over 10 million advertisers
  - A Facebook user clicks on 12 ads on average every month
- On average, users spend 34 minutes on Facebook every day
- There were over 3.5 billion live feeds on Facebook towards the end of 2018
- 500 million people use Facebook Stories daily

https://www.omnicoreagency.com/facebook-statistics/

NoSQL Databases

Five Advantages

1. Elastic scaling

- “Classical” database administrators scale up – buy bigger servers as database load increases
- Scaling out – distributing the database across multiple hosts as load increases

2. Big Data

- Volumes of data that are being stored have increased massively
- Opens new dimensions that cannot be handled with RDBMS

http://www.techrepublic.com/blog/10things/10-things-you-should-know-about-nosql-databases/1772

NoSQL Databases

Five Advantages

3. Goodbye DBAs (see you later?)

- Automatic repair, distribution, tuning, … vs. expensive, highly trained DBAs of RDBMSs

4. Economics

- Based on cheap commodity servers → less costs per transaction/second

5. Flexible Data Models

- Non-existing/relaxed data schema → structural changes cause no overhead

NoSQL Databases

Five Challenges

Less and less critical

1. Maturity

- Still in pre-production phase
- Key features yet to be implemented

2. Support

- Mostly open source, result from start-ups
  - Enables fast development
- Limited resources or credibility

3. Administration

- Require lot of skill to install and effort to maintain

NoSQL Databases

Five Challenges

4. Analytics and Business Intelligence

- Focused on web apps scenarios
  - Modern Web 2.0 applications
  - Insert-read-update-delete
- Limited ad-hoc querying
  - Even a simple query requires significant programming expertise

5. Expertise

- Few number of NoSQL experts available in the market

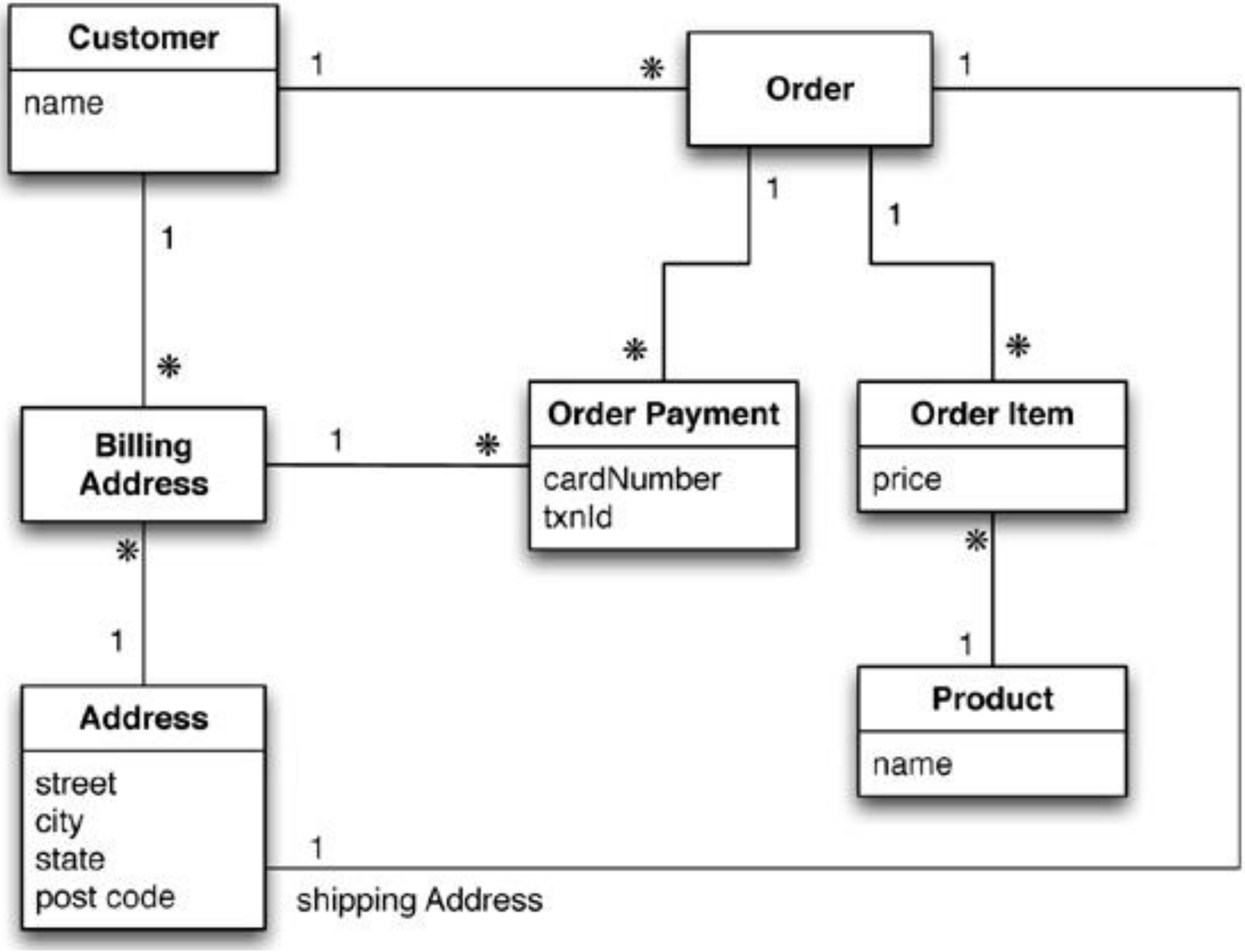
Data Assumptions

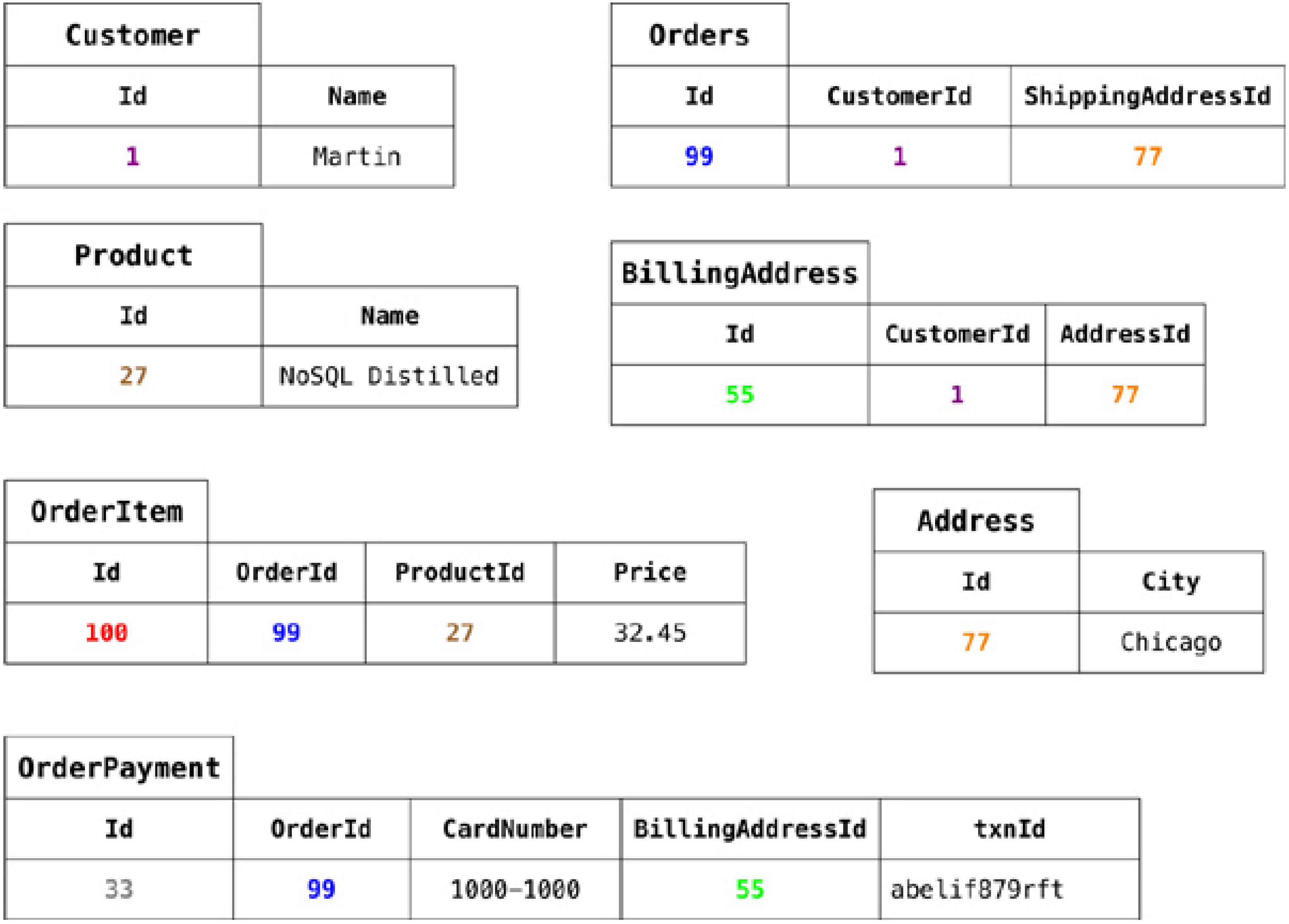
RDBMS	NoSQL
integrity is mission-critical	OK as long as most data is correct
data format consistent, well-defined	data format unknown or inconsistent
data is of long-term value	data are expected to be replaced
data updates are frequent	write-once, read multiple (no updates, or at least not often)
predictable, linear growth	unpredictable growth (exponential)
non-programmers writing queries	only programmers writing queries
regular backup	replication
access through master server	sharding across multiple nodes

NoSQL Data Model

Aggregates

- Data model = the model by which the database organizes data
- Each NoSQL solution has a different model
  - Key-value, document, column-family, graph
  - First three orient on aggregates
- Aggregate
  - A data unit with a complex structure
    - Not just a set of tuples like in RDBMS
  - Domain-Driven Design: “an aggregate is a collection of related objects that we wish to treat as a unit”
    - A unit for data manipulation and management of consistency


Customer
name
Order
Billing Address
Order Payment
cardNumber
txnId
Order Item
price
Address
street
city
state
post code
Product
name
shipping Address
1
*
1
1
1
*
*
1
*
*
1
*
1
*
1
1



Customer

Id	Name
1	Martin

Orders

Id	CustomerId	ShippingAddressId
99	1	77

Product

Id	Name
27	NoSQL Distilled

BillingAddress

Id	CustomerId	AddressId
55	1	77

OrderItem

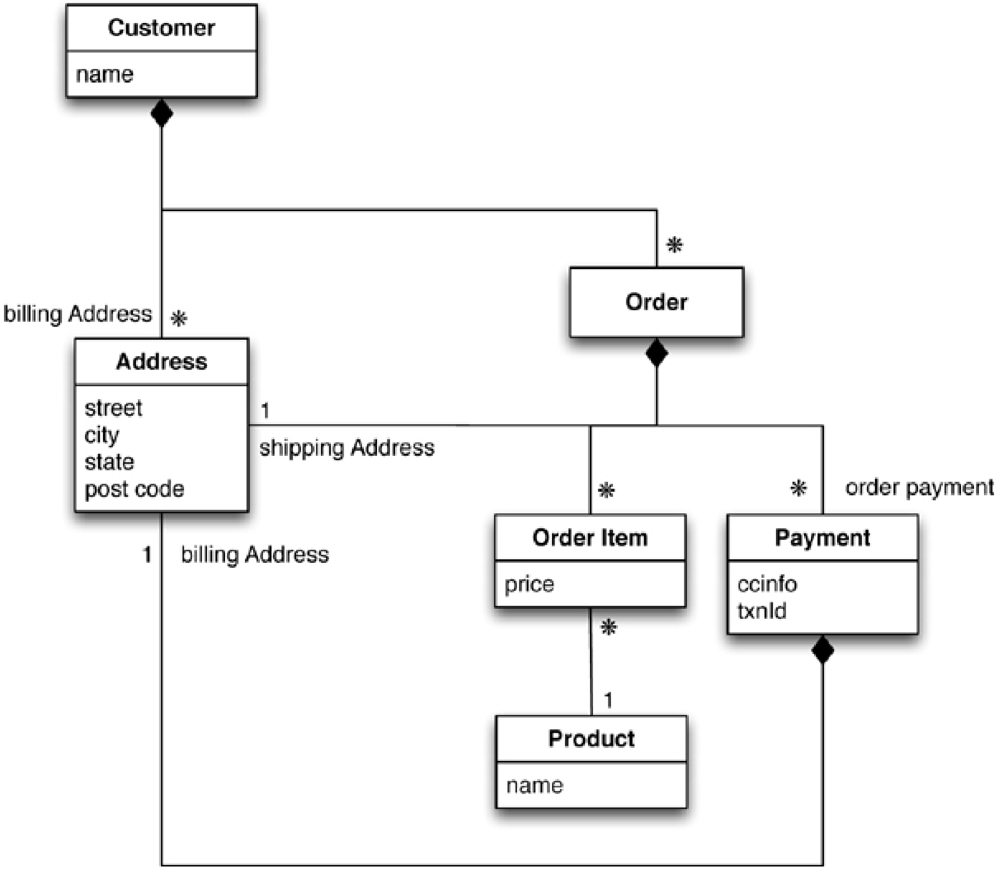
Id	OrderId	ProductId	Price
100	99	27	32.45

Address

Id	City
77	Chicago

OrderPayment

Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft

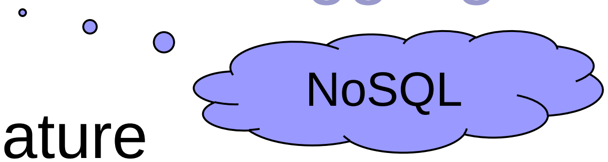


```
// in customers
{
"customer": {
"id": 1,
"name": "Martin",
"billingAddress": [{"city": "Chicago"}],
"orders": [
  {
    "id":99,
    "customerId":1,
    "orderItems":[
    {
    "productId":27,
    "price": 32.45,
    "productName": "NoSQL Distilled"
    }
  ],
  "shippingAddress":[{"city":"Chicago"}]
  "orderPayment":[
    {
    "ccinfo":"1000-1000-1000-1000",
    "txnId":"abelif879rft",
    "billingAddress": {"city": "Chicago"}
    }],
  }]
}
}
```

NoSQL Data Model

Aggregates – aggregate-ignorant

- There is no universal strategy how to draw aggregate boundaries
  - Depends on how we manipulate the data
- RDBMS and graph databases are aggregate-ignorant
  - It is not a bad thing, it is a feature
  - Allows to easily look at the data in different ways
  - Better choice when we do not have a primary structure for manipulating data



NoSQL Data Model

Aggregates – aggregate-oriented

- Aggregate orientation
  - Aggregates give the database information about which bits of data will be manipulated together
    - Which should live on the same node
  - Helps greatly with running on a cluster
    - We need to minimize the number of nodes we need to query when we are gathering data
- Consequence for transactions
  - NoSQL databases support atomic manipulation of a single aggregate at a time

NoSQL Databases

Materialized Views

- Disadvantage: the aggregated structure is given, other types of aggregations cannot be done easily
  - RDBMSs lack of aggregate structure → support for accessing data in different ways (using views)
- Solution: materialized views
  - Pre-computed and cached queries
- Strategies:
  - Update materialized view when we update the base data
    - For more frequent reads of the view than writes
  - Run batch jobs to update the materialized views at regular intervals

NoSQL Databases

Schemalessness

- When we want to store data in a RDBMS, we need to define a schema
- Advocates of schemalessness rejoice in freedom and flexibility
  - Allows to easily change your data storage as we learn more about the project
  - Easier to deal with non-uniform data
- Fact: there is usually an implicit schema present
  - The program working with the data must know its structure

References

- http://nosql-database.org/
- Pramod J. Sadalage – Martin Fowler: **NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence**
- Eric Redmond – Jim R. Wilson: **Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement**
- Sherif Sakr – Eric Pardede: **Graph Data Management: Techniques and Applications**
- Shashank Tiwari: **Professional NoSQL**